New York Traffic Data Investigations

By Brian Finnerty

Introduction

To cover the point of this paper and coding exercise, I am delving deeper into the topic of Data Science as a personal project for myself. Many aspects of life were impacted by the COVID-19 pandemic that struck during the early months of 2020. This is especially true for New York which over the course of the year held a travel ban that restricted traffic in and out of the state and it required travelers that did visit the state to have a fourteen day quarantine period. This was only the case for individuals who spent time in states designated unsafe by New York, with some exceptions. During the height of the pandemic, almost every state in the Union was on the travel restriction list at one point or another. With many other shutdowns that ensued over the course of the pandemic, travel in general was ground to a halt.

With many interesting travel examples to look at, I chose to focus in on New York due to the particular travel ban it held and because of the troubles the ban posed to myself during last year. This paper will initially look at how the pandemic affected speeds of traffic over the course of 2015-2020 but later in later iterations of this project I aim to expand the scope of my data analysis.

Data Cleaning

With millions of rows of data entries, much of the cleaning was performed through various python scripts, knowledge of the data set, and MySQL queries on the relational database I generated. Although the majority of the data was very clean when scraped from the New York Department of Transportation's public cvs, certain columns for the short count speed of the HDSB Raw Traffic Data set were unnecessary or left unfilled. Some of these included the entirely empty columns like Latitude, Longitude, Unclassified, and Flag Field. Unfortunately, these columns were left with nulls which would have otherwise provided a richer investigation into changing traffic patterns.

Other various data cleaning techniques like applying One-Rules to various null data or impossible values were utilized to provide a more comprehensive analysis on the data set.

Data Processing and Tools

Much of the data processing for this project was performed through Python and various packages like Pandas, Numpy and MatPlotLib to manipulate, clean, standardize, and visualize the data. All of the data was hosted on a local MySQL database that would allow for it to be loaded into the scripts and manipulated while preserving the structure of the original data. Although not all data cleaning techniques are captured in the codebase on this Github project, these tools were integral to performing investigations on the numerous raw traffic data CSV files pulled from the NY Department of Transportation.

Investigations

This section will cover the different investigations performed on the data, with room for extension in the future. Currently, this project only covers the changes of traffic speeds over the course of 2015 - 2020 by looking at speeding vehicles vs slow vehicles. Speeding vehicles are defined as ones that exceed 5 MPH over the designated speed limit for the road. Although some may claim that exceeding the speed limit at all can be considered speeding, many Americans and law enforcement do not hold that to be the true standard. Any excessive speed that is typically punished with a ticket is considered speeding. The following graphs will look into the various administrative traffic regions New York state is divided into. Refer to Figure 1 for further detail on these regions.

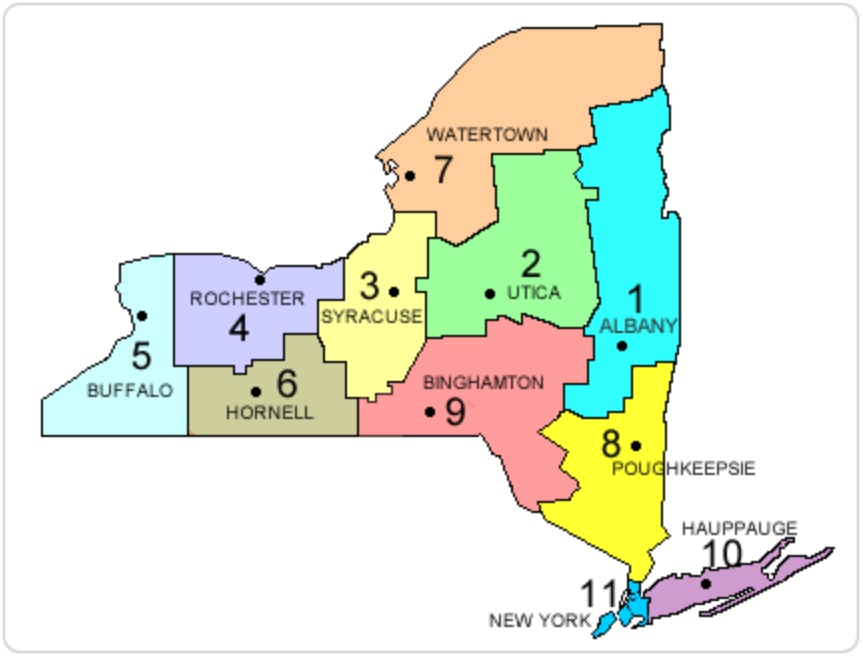
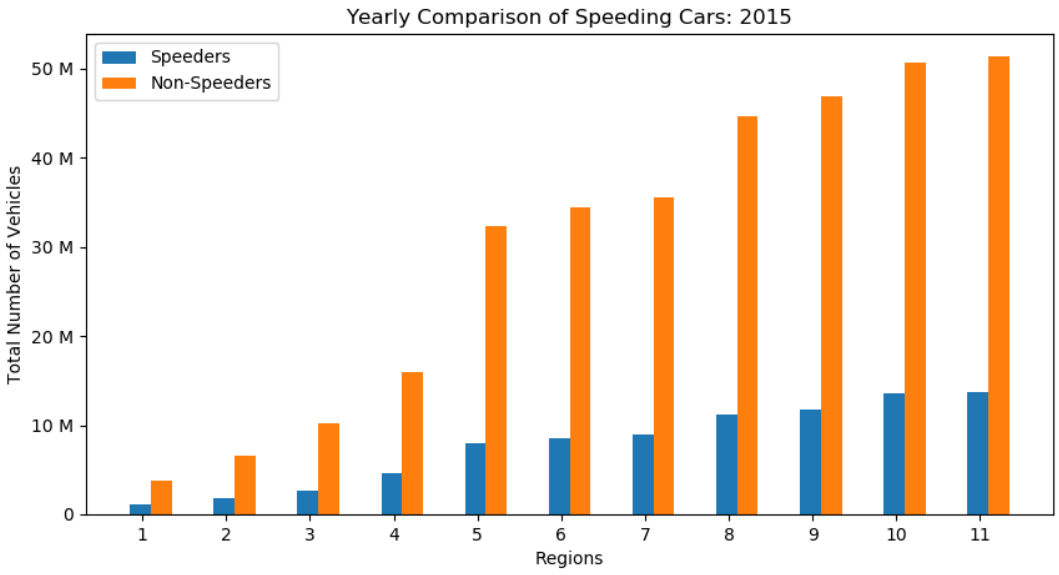
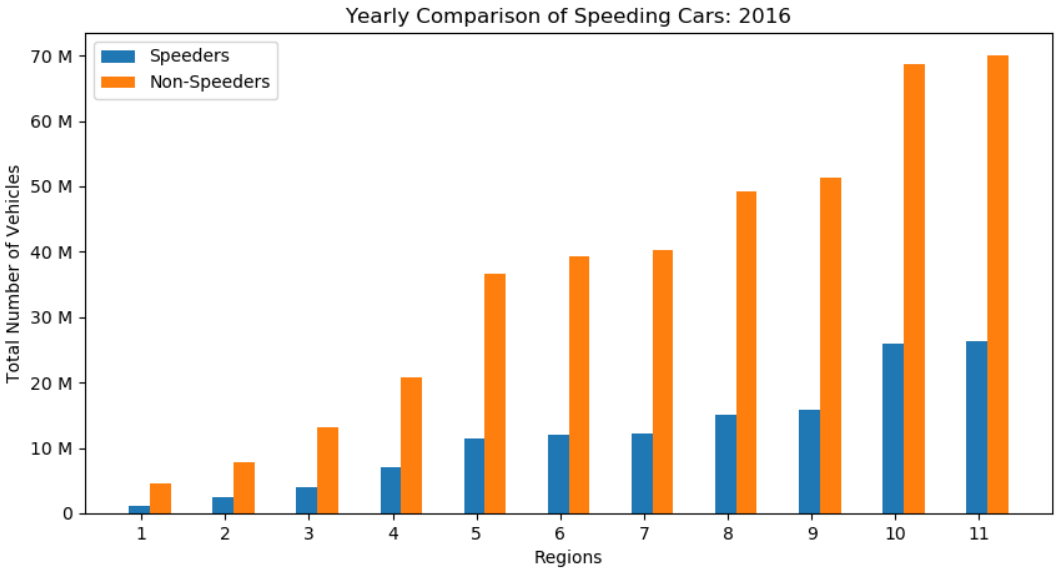
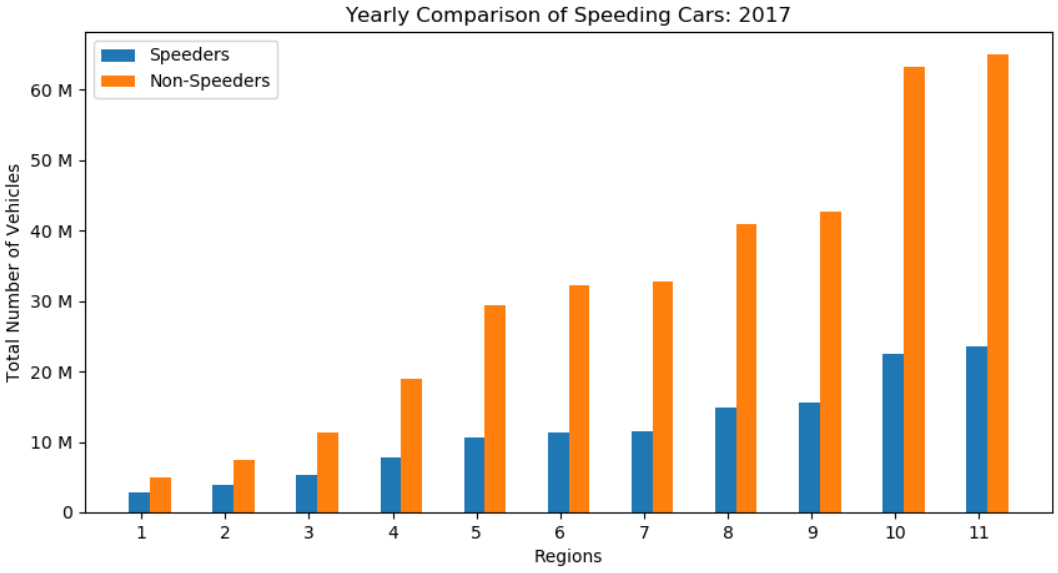


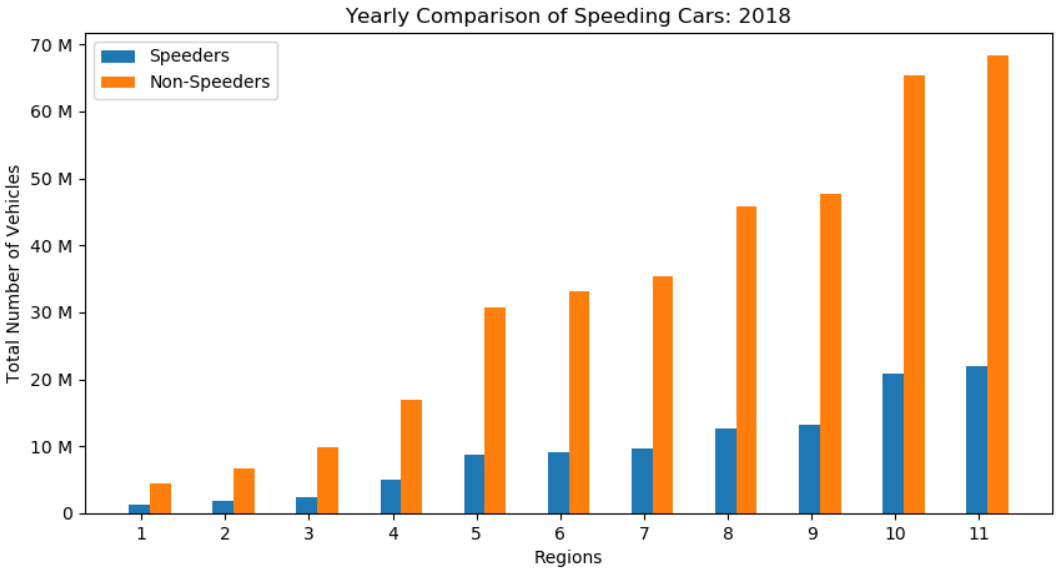
Figure 1: New York State's Traffic Regions managed by the DOT

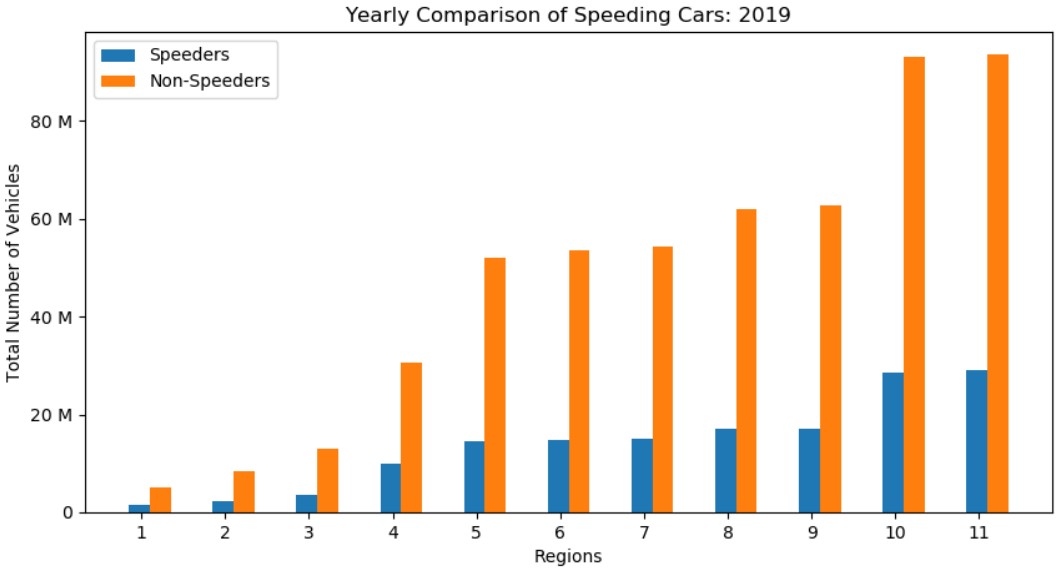
1. Changes in Regional speeds 2015-2020

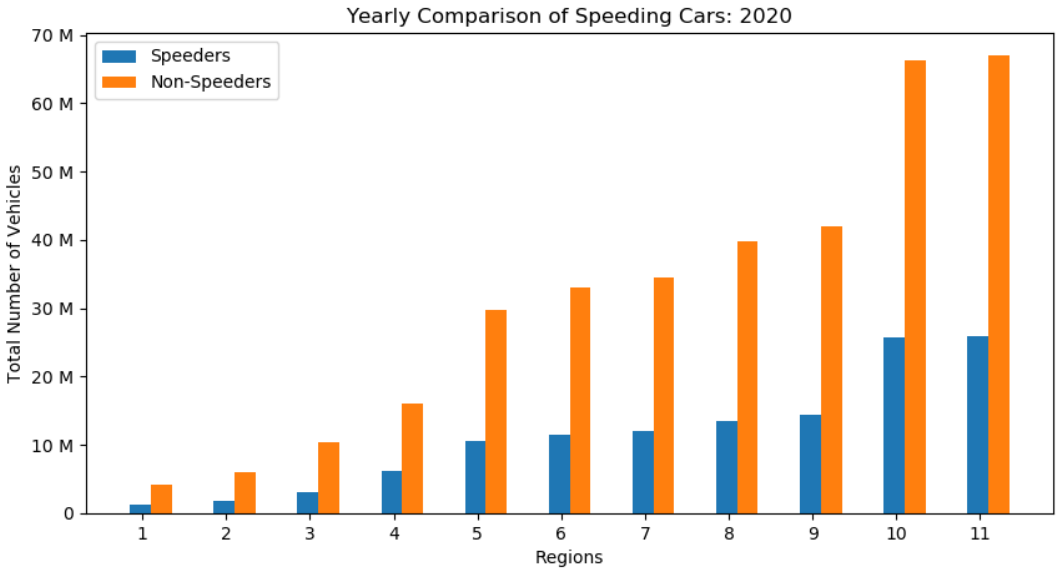



Yearly Comparison of Speeding Cars: 2016
Speeders
Non-Speeders
Total Number of Vehicles
70 M
60 M
50 M
40 M
30 M
20 M
10 M
0
1
2
3
4
5
6
7
8
9
10
11
Regions


Yearly Comparison of Speeding Cars: 2017
Speeders
Non-Speeders
Total Number of Vehicles
60 M
50 M
40 M
30 M
20 M
10 M
0
1
2
3
4
5
6
7
8
9
10
11
Regions


Yearly Comparison of Speeding Cars: 2018
Speeders
Non-Speeders
70 M
60 M
50 M
40 M
30 M
20 M
10 M
0
Total Number of Vehicles
1
2
3
4
5
6
7
8
9
10
11
Regions


Yearly Comparison of Speeding Cars: 2019
Speeders
Non-Speeders
80 M
60 M
40 M
20 M
0
Total Number of Vehicles
1
2
3
4
5
6
7
8
9
10
11
Regions



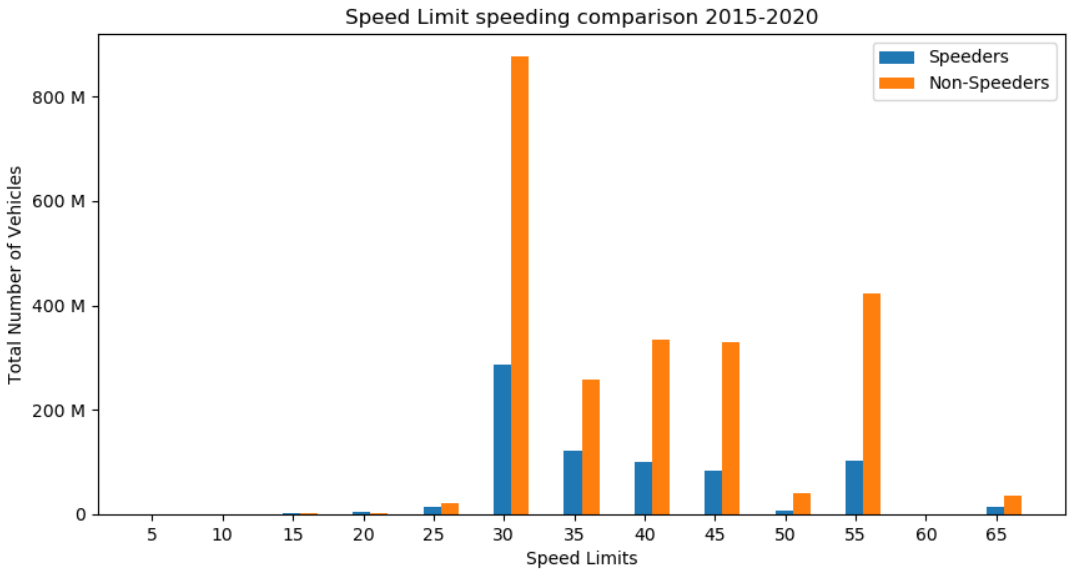
The above 5 graphs look at the relation between speeders vs non speeders in each region over the course of 2015 - 2020. One thing to note is the variable y axis that can be deceptive when looking from graph to graph. Auto Scaling the y axis was an aesthetic that allowed for the variation in lower counted regions to still shine through. Some of the most notable outcomes from these graphs are the sheer proportional difference between the cataloged amount of cars in region 10 and 11 compared to region 1 and 2. Noting Figure 1 this would make sense to those who know about population spread across where New York City is one of the most populated places in America, thus requiring two traffic regions for itself and Long Island. Of course that is not to say that these graphs will mirror population linearly as the short count speed data is dependent on when and where the monitoring stations were set up. Those setup on highways and popular roads would inevitably capture more data compared to less populous roads. Below is a table relating the proportion of Non-Speeders to Speeders in each year.

	2015	2016	2017	2018	2019	2020
Region 1	3.47	3.86	1.74	3.46	3.25	3.27
Region 2	3.71	3.12	1.93	3.67	3.57	3.26
Region 3	3.89	3.29	2.12	3.91	3.55	3.30
Region 4	3.46	2.92	2.41	3.38	3.07	2.57
Region 5	4.03	3.22	2.77	3.53	3.58	2.80
Region 6	4.01	3.28	2.84	3.65	3.60	2.86
Region 7	3.99	3.28	2.84	3.64	3.62	2.87
Region 8	3.97	3.25	2.75	3.59	3.63	2.93
Region 9	3.98	3.22	2.74	3.59	3.65	2.92
Region 10	3.74	2.64	2.81	3.14	3.25	2.57
Region 11	3.76	2.66	2.75	3.10	3.21	2.58

Table 1: Proportion of Non-Speeders to Speeders

Analysing this graph we can thankfully see that the majority of drivers on the road do not speed in New York, in fact on average for at least every 1 driver that is speeding, 3 drivers will be driving safely. The trends for speeding over the years have tended to go up after 2015 with ratios continuing to shrink with an all time low during 2020. Although these ratios do not speak to the number of crashes and deaths that occured, it can be concluded that the roadways of New York have become progressively more dangerous. These facts back up the statement of reporter David Sharp who attributed the drastically more dangerous conditions of roads in America during 2020 to “the early days of the pandemic”.
(https://apnews.com/article/covid-19-speeding-highway-deaths-30a26b82eeab5880abab5f2b30952725) He continues stating that police presence was focused on civil disobedience during the shutdown and many roadways were less congested due to travel bans and stay at home orders. An interesting extension of this view could come about with potential investigations into the volume of drivers on the roads pre and post pandemic.

2. Speeding related to Speed Limits



The above graph illustrates the proportion of speeders vs non-speeders when looking at varying speeds. For the data captured, the most common speed limit listed was 30 mph by a significant margin, with over a billion recorded instances of vehicles travelling through those areas. Although that is significant by itself, another important factor is the ratio of non-speeders to speeders as quantity itself does not indicate what speeds drivers are more likely to be found speeding.

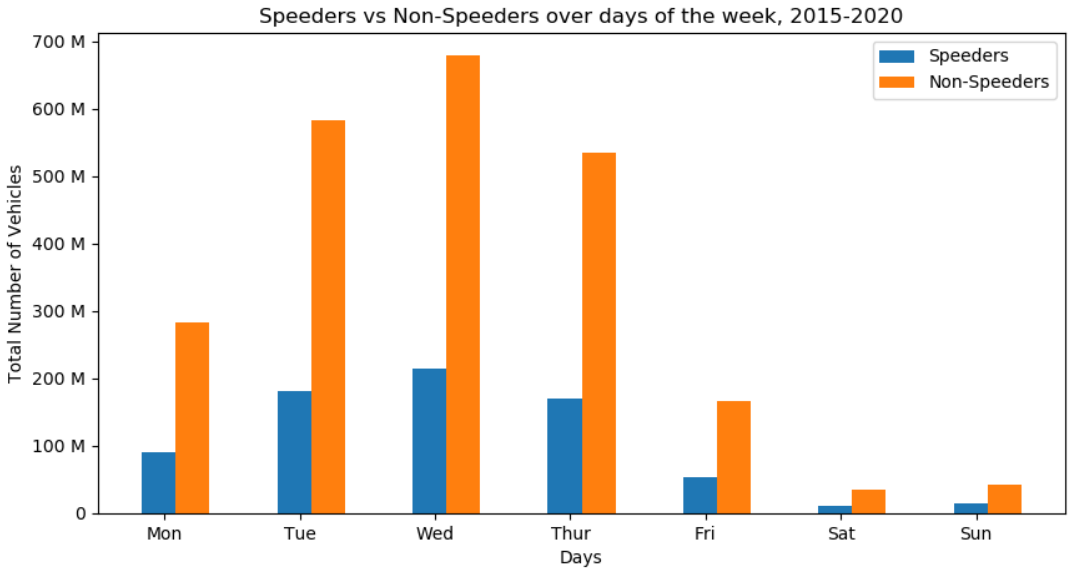
Speed Limits	Ratio
5	.02
10	20.22
15	1.75
20	.50
25	1.48

30	3.05
35	2.10
40	3.37
45	3.95
50	5.76
55	4.12
60	27.33
65	2.79

Table 2: Ration between non-speeders vs speeders categorized by speed limits

Looking at this table, there are two significant outliers with the data. Both for sixty mph and 10 mph, there is a drastic uptick in the amount of non-speeders vs speeding. With the usage of the table, it is now possible to see that at 5 mph there are many drivers willing to exceed the speed limit which may have to do with how easy it is for a vehicle to travel above 10 mph. The table also demonstrates that although 30 mph was the most commonly captured data point, the speeds of 40-55 mph are much safer for drivers. Conversely, the popular speed limits across America for school zones and urban areas of 25, 35 mph are much more likely for a driver to be speeding upon the road. Considering the vast difference between the relative frequency of each speed limit, it would suggest that most monitoring stations were placed on somewhat urban and rural roads as well as some Ny highways.

3. Speeding across Different Days



The final section covered in this paper in relation to speeding is its relation across each day. The above graph makes it immediately apparent that many of the monitoring stations were used during the middle of the work week, with very few records being logged on the weekend. However, the varying quantity of records over the course of the week does not alter the preference for speeding on these days. In fact, looking at the various speeding ratios, every day has a ratio of around 3.1. This means that each day is just as likely to have the same number of speeders, with no preference for the weekdays or weekends. Neither the bountiful amounts of free time of the weekend nor week-night rush hour draw more speeders than the other. This consistency however does not preclude any day from being more dangerous as this data set merely identifies speeding tendency rather than crashes.

Conclusion

The variety of investigations that I have performed on the dataset so far has only brought about more questions that I wish to have answered. While every day has a consistent tendency for speeding, the different speed limits in New York have a wide difference for speeders and non-speeders. In addition to this, these investigations bring up the startling reality that speeding was much more commonplace during 2020.

Although the particular reasons can not fully be concluded at the moment by this paper, my initial hypothesis would be one of less traffic on the roads. Further improvements to this paper would be to take a look at other datasets provided by the New York Department of Transportation in relation to the continuous count of both volume and vehicle classification to see how the types of cars have changed over the year as well as congestion on the roads. Ideally, these next steps will paint a much clearer picture onto how Covid-19 affected traffic in New York State.